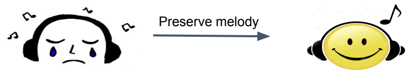
# Music Sentiment Transfer

Michael Zhou, Miles Sigel, Jiebo Luo

## Abstract

Audio sentiment transfer is an **extension of the style transfer task, a popular application of machine learning on media data**. Originally studied on images, style transfer is meant to iteratively approach an equilibrium between style and content characteristics, changing the style of a source image to that of a target. Moreover, **sentiment transfer is a natural evolution of style transfer, as sentiment transfer is rooted in applying the sentiment of a source to be the new sentiment for a target piece of media.** First introduced in this paper, music style transfer attempts to apply the high level objective of sentiment transfer to the domain of music. We propose a **CycleGAN approach to bridge disparate domains.** In order to use the network, we choose to use symbolic, MIDI, data as the music format. Through the use of a cycle consistency loss, we are able to create injective mappings that preserve the content and realism of the source data while transferring sentiment. **Results and literature suggest that the task of music sentiment transfer is more difficult than image sentiment transfer because of the temporal characteristics of music and lack of existing datasets.**


Preserve melody

## Introduction

Style transfer is a topic in computer vision pertaining to **transferring the artistic styletyle from one media to another.** This task, originally studied in images and introduced by Gaytes et al., applied the style characteristics (ex. color, texture, brush strokes) of a source image to a target image through an iterative process that minimized the distance between "style characteristics." Sentiment transfer is another task similar to style transfer; however, sentiment transfer focuses on **a higher level aspect, sentiment, which modifies perceptual features that don't directly distort the identity of the image**, something that existing style transfer cannot accomplish.

Style transfer in audio takes form based on the different types of audio tasks being studied. Voice conversion networks are used to change a raw audio voice data from male to female. Yet the field of **audio style transfer is greatly limited by the lack of paired data**, as general frequency conversion is possible but a fine grained mapping is difficult. Most hindered by such a problem is that of this paper, music style transfer. The domain of music style transfer brings into question the foundation of style in music. Contrary to image style transfer with clear pixel level style characteristics (color, brightness), music incorporates the time-domain, different instrumentation, and musical clef differencers. As a consequence of such limitations of high-resolution raw audio, **audio research has gravitated toward the symbolic form of music commonly used today, MIDI** (Musical Instrument Digital Interface).

Music sentiment transfer is an unexplored task, in which the sentiment of one piece of music is applied to another. Considering the inability of music to be separated into a clear content and style code, **domain transfer networks such as CycleGAN are common in transfer tasks involving music due to the domains being dissimilar from one another.** Brunner et al. are first to explore music genre transfer in symbolic audio. In this research, we attempt to utilize the genre transfer network on a novel dataset containing binary sentiment labels. **In this task, we attempt to transfer music from a negative sentiment domain to a positive sentiment domain and vice versa.**

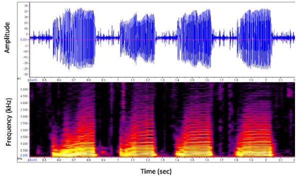
## Data Representation

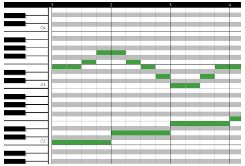
### Data-Driven vs Knowledge-Driven Approach

| Data-Driven | Knowledge-Driven |
|---|---|
| **Machine learning** to learn the **underlying patterns of music sentiments** from data. | **Music theory** to change **musical structure** (e.g. minor to major keys) |

### Music Representation: Symbolic vs Audio

| Symbolic (MIDI) | Raw Audio |
|---|---|
| Encodes **music structural properties** (key, dynamics, tone, etc.) | Common use cases: male to female voice conversion, music genre transfer<br>Unable to learn **deeper structures of music sentiments** |





## Proposed Architecture: CycleGAN

We propose a **CycleGAN** to bridge the two domains of music sentiment - happy and sad. The goal of a CycleGAN is to **learn a one-to-one mapping** between two domains **with unlabeled data**. Like other GANs, the CycleGAN consists of generators which try to generate fake data, and discriminators that tell whether the generated is real or fake. The goal is to have the generators fool the discriminator, and generate realistic images in the target domain. The CycleGAN in particular allows **transferring images back and forth between domains** while **preserving the content information**.

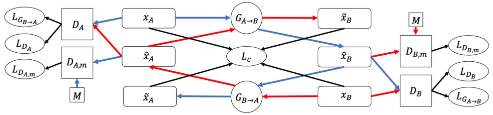
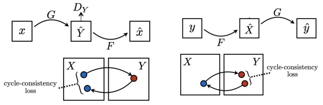


Fig. 1. Architecture of our model. The two cycles are shown in blue and red respectively. The black arrows point to the loss functions. We extend the basic CycleGAN architecture with additional discriminators $D_{A,m}$ and $D_{B,m}$.



## Discussion

**Datasets:**
The **most challenging aspect of this research is the lack of a sentiment labeled dataset**. In this experiment, we don't have access to paired data, so we needed to create our own dataset. We attempted to use existing MIDI sentiment labeled datasets, but due to the need for single track audio, we had to create our own. In order to create our own, we transformed MIDI files to their respective piano roll arrays. For future work, we suggest that a **clearly annotated dataset** be created for **both symbolic and raw audio** in a **single track**.

**Models:**
Sentiment transfer for music is different then the approach used for images due to the differing data representations. Music, once turned into discrete piano rolls, could not be separated into content and style codes. Therefore, a network like CycleGAN that uses unpaired data and can work with arbitrary data representations is an understandable choice. Autoregressive generative networks could potentially be used to accomplish this task but might be hindered by the ability to preserve source content. MelGAN-VC is another network similar to CycleGAN that uses a siamese network to preserve the distance between classes in a latent space. Models using spectrograms introduce noise, so many extant networks used for images are not congenial.

**Data Representation:**
Audio data comes in two main formats, symbolic data and raw audio data. Symbolic data is commonly represented in MIDI format, a discrete representation that annotates the notes and metadata of a piece of music. Using this format allows for less noise in the data compared to raw data, as the sampling rate of raw audio introduces many irregularities into the data. Additionally, for music, using raw data does not make some features of the underlying structure of the music apparent and relies on a network being able to learn long range temporal features. Using symbolic data, however, is potentially less applicable in real world situations with typically operate on raw audio.

## Conclusions

- Audio sentiment transfer is a **novel task** which is an extension of the better studied style transfer task.
- Audio sentiment transfer goes **beyond changing note tones**, as sentiment is embedded in **deeper structures of music**.
- There are a **lack of existing datasets** that are **labelled by sentiment for both raw audio and symbolic audio**.
- Further research should **combine an expert knowledge** of musical composition to tune parameters and **provide insight into the relationship between musical structure and sentiment**.

**Contact:**

Michael Zhou
Cornell University
mgz27@cornell.edu
425-324-6122

Miles Sigel
Rice University
milesigel@gmail.com
512-987-3780

**References:**

1. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.midi.org%2Fforum%2F808-midi-file-into-sheet-music&psig=AOvVaw3CX0xhvidb8OYdxZR6hqUr&ust=1626885668369000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCLDpiseL8vECFQAAAAAdAAAAABAJ
2. Convenient Fish Acoustic Data Collection in the Digital Age - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Spectrograms-and-Oscillograms-This-is-an-oscillogram-and-spectrogram-of-the-boatwhistle_fig2_267827408 [accessed 20 Jul, 2021]
3. CycleGAN Blog https://hardikbansal.github.io/CycleGANBlog/
4. J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244. https://arxiv.org/pdf/1703.10593.pdf
5. Matthew Amodio, & Smita Krishnaswamy (2019). TraVeLGAN: Image-To-Image Translation by Transformation Vector Learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8975-8984. https://arxiv.org/abs/1703.10593
6. Gatys, L., Ecker, A., & Bethge, M. (2016). A Neural Algorithm of Artistic Style. *Journal of Vision*, *16*(12), 326. https://doi.org/10.1167/16.12.326
7. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y., & Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. *Advances in Neural Information Processing Systems*, *32*(NeurIPS 2019). https://papers.nips.cc/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf